# Pre-class activities
# Naive Bayes Classification

Emmanuel Rachelson

## 1 Cluedo

British people write "rigour" while US citizens write "rigor". A suspect in a hotel wrote this word on a piece of paper, then burnt it and left it behind to be found by the police. Only one letter remains readable; it is a vowel. It is known that 40% of the hotel's clients are British and 60% are Americans.
☞ What is the probability the suspect is British?

*By using Bayes formula, one obtains that the probability of the suspect being British is equal to the likelihood of finding a vowel in the word provided that a British citizen wrote it, times the probability of meeting a British citizen in this hotel, divided by the probability of finding a vowel in this word in general.*

$$\begin{aligned}\mathbb{P}(British|vowel) &= \frac{\mathbb{P}(vowel|British)\mathbb{P}(British)}{\mathbb{P}(vowel)} \\ &= \frac{\mathbb{P}(vowel|British)\mathbb{P}(British)}{\mathbb{P}(vowel|British)\mathbb{P}(British) + \mathbb{P}(vowel|US)\mathbb{P}(US)} \\ &= \frac{\frac{3}{6}\frac{40}{100}}{\frac{3}{6}\frac{40}{100} + \frac{2}{5}\frac{60}{100}} \\ &= \frac{5}{11}\end{aligned}$$

☞ Suppose now there are more than two english-speaking nationalities staying in this hotel. What part of the previous calculation is unecessary to establish the most probable citizenship of the suspect?

*To establish the most probable citizenship of the suspect, one needs to compare* $\mathbb{P}(country|vowel)$ *for all values of country and pick the one with the highest estimated probability. Between all these calculations, only the numerator changes, hence:*

$$most\ probable\ citizenship = \underset{country}{\operatorname{argmax}}\, \mathbb{P}(vowel|country)\mathbb{P}(country)$$

## 2 [Optional] Quality check

In a manufacturing line, 1% of the production has a defect. A performance test allows to filter out 95% of faulty products but also excludes 2% of acceptable ones.

☞ What is the probability of a control error?

*The probability of a control error is the probability of either a false positive (FP) or a false negative (FN). Let us consider the events "faulty product" (F) and "rejected product" (R). The data tells us that:*

$$\begin{aligned}\mathbb{P}(F) &= 0.01\\ \mathbb{P}(R|F) &= 0.95\\ \mathbb{P}(R|\bar{F}) &= 0.02\end{aligned}$$

*Then:*

$$\begin{aligned}\mathbb{P}(FN) &= \mathbb{P}(F \wedge \bar{R}) = \mathbb{P}(\bar{R}|F)\mathbb{P}(F) = (1 - \mathbb{P}(R|F))\,\mathbb{P}(F)\\ \mathbb{P}(FP) &= \mathbb{P}(\bar{F} \wedge R) = \mathbb{P}(R|\bar{F})\mathbb{P}(\bar{F})\end{aligned}$$

*Thus:*

$$\mathbb{P}(FN \vee FP) = (1 - 0.95)\,0.01 + 0.02\,(1 - 0.01) = 0.0005 + 0.0198 = 0.0203$$

*Interestingly, the mot probable control errors concern acceptable products rejected by mistake.*

☞ What is the probability of a letting a faulty product go through?

*This probability is $\mathbb{P}(F|\bar{R})$. By application of Bayes' theorem:*

$$\begin{aligned}\mathbb{P}(F|\bar{R}) &= \frac{\mathbb{P}(\bar{R}|F)\mathbb{P}(F)}{\mathbb{P}(\bar{R})}\\ &= \frac{(1 - \mathbb{P}(R|F))\,\mathbb{P}(F)}{\mathbb{P}(\bar{R}|F)\mathbb{P}(F) + \mathbb{P}(\bar{R}|\bar{F})\mathbb{P}(\bar{F})}\\ &= \frac{(1 - \mathbb{P}(R|F))\,\mathbb{P}(F)}{(1 - \mathbb{P}(R|F))\,\mathbb{P}(F) + \left(1 - \mathbb{P}(R|\bar{F})\right)(1 - \mathbb{P}(F))}\\ &= \frac{(1 - 0.95)0.01}{(1 - 0.95)0.01 + (1 - 0.02)(1 - 0.01)}\\ &= 0,000515092\end{aligned}$$